Interdiscip Sci Comput Life Sci (2017) 9:36–45
DOI 10.1007/s12539-017-0214-y

ORIGINAL RESEARCH ARTICLE

Analysing Algorithms and Data Sources for the Tissue-Specific Reconstruction of Liver Healthy and Cancer Cells

Jorge Ferreira[1] · **Sara Correia**[1] · **Miguel Rocha**[1]

Received: 21 July 2016 / Revised: 12 December 2016 / Accepted: 2 January 2017 / Published online: 2 March 2017

Abstract Genome-Scale Metabolic Models (GSMMs), mathematical representations of the cell metabolism in different organisms including humans, are resourceful tools to simulate metabolic phenotypes and understand associated diseases, such as obesity, diabetes and cancer. In the last years, different algorithms have been developed to generate tissue-specific metabolic models that simulate different phenotypes for distinct cell types. Hepatocyte cells are one of the main sites of metabolic conversions, mainly due to their diverse physiological functions. Most of the liver's tissue is formed by hepatocytes, being one of the largest and most important organs regarding its biological functions. Hepatocellular carcinoma is, also, one of the most important human cancers with high mortality rates. In this study, we will analyze four different algorithms (MBA, mCADRE, tINIT and FASTCORE) for tissue-specific model reconstruction, based on a template model and two types of data sources: transcriptomics and proteomics. These methods will be applied to the reconstruction of metabolic models for hepatocyte cells and HepG2 cancer cell line. The models will be analyzed and compared under different perspectives, emphasizing their functional analysis considering a set of metabolic liver tasks. The results show that there is no "ideal" algorithm. However, with the current analysis, we were able to retrieve knowledge about the metabolism of the liver.

Keywords Tissue-specific genome-scale metabolic models · Liver metabolism · Hepatocellular carcinoma

✉ Miguel Rocha
mrocha@di.uminho.pt

[1] Centre Biological Engineering, Universidade do Minho, Campus de Gualtar, 4710-057 Braga, Portugal

1 Introduction

As one of the most important tools to investigate cell metabolism, Genome-Scale Metabolic Models (GSMMs) comprise a mathematical formulation of the biochemical reactions' network of a given organism [1]. They are based on the assumption of a pseudo-steady state and the use of a stoichiometric matrix to be able to perform simulations using numerical optimization [2]. In addition, GSMMs can simulate phenotypes under different conditions (genetic, physicochemical and environmental) that are imported to the model in the form of constraints, taken into account when the optimization is being performed for the prediction of fluxes [3].

The decrease in the cost of high-throughput *omics* data and the scientific advances in bioinformatics have enabled metabolic reconstructions, not only for smaller organisms, but also for eukaryotes and even humans. The first metabolic human GSMM was released in 2007, Recon1, with a very thorough manual curation allowing to validate each single metabolic reaction [4]. This revolutionized the development of new algorithms specifically made for human models. A small set of modifications on the normal metabolism may imply serious consequences. Diseases such as obesity, diabetes, cancer are some of the examples [5]. GSMMs provide important tools to study these disturbances.

Today, the most extensive human metabolic models are Recon2 [6] and HMR 2.0 [7]. The Recon2 is the result of the merge of other metabolic components present in Recon1 together with the information from Edinburgh Human Metabolic Network (EHMN) [8], HepatoNet1 [9], a module containing information about acyl carnitine and fatty acid oxidation [10], and another model with data about the human small intestinal enterocyte

[11]. Regarding the HMR 2.0, it also contains integrated information from the Recon1, EHMN, HepatoNet1, iHuman1512 [12] and iAdipocytes1809 [13], and also information from the four major metabolic databases (KEGG [14], HumanCyc [15], LIPID MAPS Lipidomics Gateway [16] and REACTOME [17]).

Due to the evolution of the omics technologies, several types of data can be used in combination with the models to improve their predictive power [18, 19], to fill gaps of knowledge. Although the analysis of fluxes (fluxomics) seems the better approach to complete the models, it can only be used to determine a specific set of reactions. Metabolomics would also be a good choice, but due to the fact that a metabolite takes part of multiple reactions, its measurement is not in many cases biologically relevant.

With this in mind, proteomics and transcriptomics data make possible the analysis of the current molecular state of the organism. Although proteomics is not as advanced as transcriptomics, the Human Protein Atlas (HPA) contains data obtained by immunostaining experiments associated with their subcellular localization [3].

Transcriptomics data (mRNA) can be measured with more precision and wider range when compared to proteins, with more automated processes and lower costs associated. Although these data are more frequently used, the acquisition of knowledge is more difficult due to the fact that there are different layers to take into account, like the translation, post-translational modifications, mRNA/protein degradation or enzyme regulation by activators or inhibitors [20–22].

Metabolic models, when associated with information on environmental conditions (e.g., growth medium), can provide a reasonable prediction of several metabolic phenotypes, such as growth rates, nutrient uptake rates, compound excretion rates or gene essentially [23].

Flux Balance Analysis (FBA) is the most used technique to predict the phenotype using a model [24]. Based on the assumption of a pseudo-steady state, it has been accepted as one of the most robust methods to study the physiology of the cell (given a set of constrains to the model). This assumption implies that all the internal metabolites are "balanced" and the cell has a tendency to optimize a defined objective function (the usual one is the maximization of the cell growth).

Although these models are simply a mathematical representation of a cell, they have proved that their application can have an high value for biomedical purposes. Characteristics like its ease of implementation or their predictive power have made possible the prediction of which genes to manipulate in metabolic engineering (production of shikimic acid and putrescine in *E. coli*) [25], predict drug targets (essential metabolites considered critical to the *Vibrio vulnificus* CMCP6) [26], and specific cells linked to diseases, for example, hepatocytes from patients who suffered nonalcoholic fatty liver disease [7].

GSMMs were also used to simulate cancer cells' metabolism and address drug target discovery, to study tumor suppressors and oxidative stress [12, 27–29]. One study on the fumarate hydratase (FH) enzyme (from the TCA cycle) with a mutation that led to the loss of its function and had diseases associated with it, like renal cancer, and on the heme pathway, demonstrated that "wild-type" cells are not affected when targeting the cancer cells with a drug to a specific target [27].

If we take a deep look into the complex human organism, we have several types of cells in different proportions and each of them with distinct roles in the metabolism. For this and other reasons, it is of crucial importance the "creation" of tissue-specific metabolic models that can lead to an improvement of metabolic phenotypes and their related diseases. Indeed, the creation of methods for the integration of omics data for the generation of tissue/cell type-specific metabolic models is of crucial significance for a better understanding of the biochemical and genetic complexity of the human metabolism [30]. Establishing several models that can simulate diverse cell types from human tissues may be a good starting point for a better understanding of complex diseases [12].

Liver is one of the most crucial human organs regarding the metabolism, being responsible for the removal of toxic substances and regulation of the bile, plasma and red cells [5]. Two different types of cells comprise the liver, the parenchymal and nonparenchymal cells and the most common diseases associated are hepatitis, hepatocellular carcinoma (HCC) and nonalcoholic fatty liver disease (which can all be linked to disorders in the metabolism) [31].

HCC affects many humans in the world with half a million new cases per year [32]. Given the fact that there is a huge amount of data produced by high-throughput technologies (and the continuous decrease on its costs), the understanding of the main dissimilarities between the healthy and disease conditions can elucidate the underlying mechanism of the liver cells and their related diseases [33].

To unveil the mechanisms behind the metabolism of the liver, several tissue-specific metabolic model reconstruction algorithms have been utilized for hepatocytes [34, 35] and even a manually curated one, the HepanoNet1 [9]. In previous work, a systematic analysis of different algorithms for the same purpose was conducted by the authors [36].

Here, the objective is to extend the previous work, by taking into account the normal hepatocytes and also reconstructing a model for the HepG2, a liver cancer cell line [37], also increasing the number of algorithms tested to consider the most recent ones. For both conditions, we will consider different data sources for transcriptomics and proteomics. We aim to analyze the models generated for

both conditions and compare how the algorithms and data sources affect their functional and structural capabilities, achieving more knowledge on how different is the metabolism between the studied conditions, highlighting the reactions or pathways affected.

2 Materials and Methods

2.1 Models and Data

Recon1 will be used as the GSMM, which comprises 3742 reactions, 2766 metabolites, 2004 proteins and 1905 genes [38]. Proteomics data were retrieved from the Human Protein Atlas (HPA) [39], which contains information on protein concentration levels. Here, we used HPA data (version 14) for the HepG2 cell line derived from a hepatocellular carcinoma [37] and hepatocytes from normal liver tissue data.

For the transcriptomics data, we used raw expression data from 3 different samples of HepG2 cell lines (from GSE7307 dataset from Gene Expression Omnibus (GEO)) and the information present in the Gene Expression Barcode (GEB) [40] for the hepatocytes from normal liver tissue data. The processing of the raw data used was the one described in the work of Wang and his colleagues [35].

Based on them, the reaction scores were calculated using the gene–protein rules (GPRs) present on the Recon1, where the logical value "OR" will be replaced by the maximum and "AND" for the minimum of gene scores obtained by the *omics* data.

2.2 Algorithms for Tissue-Specific Model Reconstruction

Given the different algorithms to create tissue-specific metabolic models based on a generic human model, we briefly explain the four that will be used in this work. The pseudo code of all the algorithms described can be viewed in Table 1.

2.2.1 INIT/tINIT

The Integrative Network Interface for Tissues (INIT) algorithm maximizes the matches between reaction states (active or inactive) and data regarding expression or non-expression of genes/proteins, returning flux values and a tissue-specific model (i.e., a set of reactions from the original model considered to exist in the tissue). The method solves a Mixed Integer Linear Program (MILP), where binary variables represent the presence of each reaction from the template model in the resulting model. Although the algorithm normally uses proteomic data from HPA, transcriptomics can also be given as an input.

In the definition of the objective function, positive weights are given to reactions with a higher evidence from the input, and negative to the ones who have low or no expression. If there is supportive information (usually metabolomics) that corroborate the presence of a certain metabolite, the necessary reactions may be included to the final model to produce it [12]. The task-driven INIT (tINIT) is an extension of the previous algorithm [41]. The improvement is based on the possibility to define a metabolic task in agreement with the context of the reconstruction. These may be the consumption or production of a metabolite or activation of the reactions of a particular pathway for the tissue.

2.2.2 MBA

Differently from INIT, the Model-Build Algorithm (MBA) [34] returns only a final model and no flux values. This algorithm accepts as input a generic metabolic model and two sets of reactions. The first set (C_H) comprises reactions with high support (e.g., literature) for the inclusion on the final model, while the other one (C_M) contains usually information derived from high-throughput data. In an iterative way, all the non-core reactions (based on the previously established sets) are removed in a random order, while the model is tested for consistency. The iteration ends when all the reactions have been submitted for the removal test in the final model. The final model should contain the whole set of the C_H, a maximum number of reactions from the C_M and the least possible of the remaining non-core ones, normally requested to avoid connectivity issues.

Since the order by which each reaction is tested for removal matters, there is the need to repeat this algorithm several times to obtain a set of models. The final one should be a model based on the ranking of the frequency of the reactions in the set, adding them to the C_H core until a coherent model is found [34].

2.2.3 mCADRE

The Metabolic Context specificity Assessed by Deterministic Reaction Evaluation (mCADRE) [35] algorithm is quite similar to the MBA, but only requires the reconstruction of a single model. It is initialized by ranking the reactions on the original model using three distinct scores: confidence, expression and connectivity. With the help of a threshold value for the scores, a core of reactions and the order of removal of the non-core ones is established.

The input for the algorithm considers the frequency of expression states in a set of profiles (requiring a change of the data to binary values), instead of levels of expression.

Table 1 Formulation and description of algorithms of MBA, tINIT, mCADRE and FASTCORE

MBA	tINIT
$generateModel(R_G, C_H, C_M)$ $R_P \leftarrow R_G$ $R_S \leftarrow R_P \backslash (C_H \cup C_M)$ $P \leftarrow randomPermutation(R_S)$ $for(r \in P)$ $inactiveR \leftarrow CheckModel(R_P, r)$ $e_H \leftarrow inactiveR \cap C_H$ $e_M \leftarrow inactiveR \cap C_M$ $e_X \leftarrow inactiveR \backslash (C_H \cup C_H)$ $if(\lvert e_H\rvert == 0\ AND\ \lvert e_M\rvert < \delta * \lvert e_X\rvert)$ $R_P \leftarrow R_P \backslash (e_M \cup e_X)$ $endif$ $endfor$ $returnR_P$ $endfunction$	$min \sum_{i \in R} w_i * y_i$ $s.t.$ $Sv = b$ $\lvert v_i\rvert \leq v_{max}$ $0 < v_i + (v_{max} * y_i) \leq v_{max}$ $b_j \geq \delta\ j \in Metabolomics$ $b_j = 0\ j \notin Metabolomics$ $y_{for(i)} + y_{rev(i)} \leq 1$ $v_i \geq \delta, i \in RequiredReac$ $y_i \in 0, 1$ $w_i, score\ for\ i \in R$
mCADRE	**FASTCORE**
$generateModel(R_G, treshold)$ $R_P \leftarrow R_G$ $R_C \leftarrow score(R_P) > treshold$ $coreActiveG \leftarrow flux(r)! = 0, r \in R_C$ $R_{NC} \leftarrow R_P \backslash R_C$ $for(r \in order(R_{NC}))$ $inactiveR \leftarrow CheckModel(R_P, r)$ $s1 = \lvert inactiveR \cap R_C\rvert$ $s2 = \lvert inactiveR \cap R_{NC}\rvert$ $if(r \notin withExpressionValues\ AND$ $s1 \backslash s2 <= RACIO\ AND$ $checkModelFunction(R_p \backslash inactiveR))$ $R_P \leftarrow R_P \backslash inactiveR$ $elseif(\lvert s1\rvert == 0\ AND$ $checkModelFunction(R_p \backslash inactiveR))$ $R_P \leftarrow R_P \backslash inactiveR$ $endif$ $returnR_P$ $endfunction$	$FASTCORE(N, C)$ $J \leftarrow C \cap I$ $flipped \leftarrow False, singleton \leftarrow False$ $A \leftarrow findSparseMode(J, P, singleton)$ $J \leftarrow C \backslash A$ $while(J \neq \emptyset)$ $P \leftarrow P \backslash A$ $A \leftarrow A \cup findSparseMode(J, P, singleton)$ $if(J \cap A \neq \emptyset)$ $J \leftarrow J \backslash A, flipped \leftarrow False$ $else$ $if(flipped)$ $flipped \leftarrow False, singleton \leftarrow True$ $else$ $flipped \leftarrow True$ $if(singleton)$ $\widetilde{J} \leftarrow firstElement(J)$ $else$ $\widetilde{J} \leftarrow J$ $endif$ $for(r \in \widetilde{J} \backslash I)$ $flip\ the\ sign\ in\ stoichiometric\ matrix$ $and\ swap\ the\ bounds\ of\ reaction\ r$ $endfor$ $endif$ $endif$ $endwhile$ $endFunction$

In the table, R_G represents the list of reactions from the global template model, R_C the set of core reactions on mCADRE, C_H and C_M the core and moderate probability sets used in MBA, r a reaction and the *for*(i) and the *rev*(i) represent the ith reaction direction (forward and reverse). In the FASTCORE algorithm, N is the set of all reactions in the model, C is the core set of reactions, and I the set of irreversible reactions. $J \subseteq C$ is a set with the irreversible reactions from C and $P = (N \backslash C) \backslash A$ is a "penalty" set which contains all the non-core reactions that have not been added to A

Regarding the connectivity, the reactions are ranked by the reactions in the "neighborhood". For the confidence levels, the reactions are ranked according to the evidences of that reaction in the general metabolic model.

In the process of the reconstruction, if the removal of a non-core reaction does not compromise the production of essential metabolites and the core of reactions, those reactions are removed on the previous order. However, if a particular situation requires it, the elimination of core reactions is possible.

2.2.4 FASTCORE

In a similar approach to MBA (trying not to alter the set of core reactions), the FASTCORE [42] algorithm uses another strategy by solving two Linear Problems (LP). The first maximizes the number of reactions in the core comparing the values of a reaction with a constant, while the other decreases the number of reactions that are absent in the core by minimizing the L_1-norm of the flux vector. Until the core is coherent (the whole set of core reactions is activated with the smallest number of non-core reactions), both problems are being solved alternatively and in a repeated way. For reversible reactions, the algorithm analyses both directions.

3 Results and Discussion

In this study, we reconstructed 16 models comprising all the four algorithms described above, using both conditions and data sources. Due to the fact that we are using the Recon1 metabolic model as the template model for the tissue-specific models, the data used are filtered for the genes present in this model. All the software tools used and datasets are provided, to allow for full reproducibility of the results, in a software container (using the Docker application). The image and instructions

for running it are provided in Docker Hub: https://hub.docker.com/r/saracorreia/is_cls_hepg2.

For the reconstruction of the MBA models, we generated 50 different models and merged them into a single model (the cutoffs for the creation of the core were "High" and "Medium" for the HPA data and 0.9 and 0.5 for the GEB data). For both the mCADRE and FASTCORE cores, the cutoffs were either "Medium" or 0.5, for HPA and GEB data, respectively. Finally, for the tINIT algorithm, the cutoffs for the levels "High", "Medium", "Low" and "Not Detected" were, respectively, 0.9, 0.5, 0.1 and 0. Also for the tINIT algorithm, we provided a set of specific metabolic tasks that the cell needs to perform that were given on the original paper for the algorithm [41].

For a visual understanding of the results, we display Venn Diagrams in Figs. 1 and 2, with the number of reactions that are shared for each set of conditions and different data sources.

Analysing the figures, we can tell that the algorithm that shares the most reactions between the two conditions for both data sources is the tINIT algorithm. This may be due to the nature of the algorithm because, although it cannot guarantee that the model is capable of performing all the tasks, if possible, it tries to find a set of essential reactions and ensures that those reactions in the model have flux. For the other algorithms, their percentage of shared reactions is very similar (although the MBA for HPA data has the lowest percentage of shared reactions).

The next step was to execute a hierarchical clustering process of the 16 models. With this method, we will try to identify the relations between conditions, data sources and algorithms.

Taking a closer look at the clustering results (Fig. 3), we can clearly see different patterns. Considering the three higher level clusters, the first one includes all the tINIT models (including normal and cancer cells and both data sources) and the other two all other algorithms. This can be explained by the use of the metabolic tasks in tINIT that influence heavily the set of reactions to be included in all models. A further analysis could be to try to find alternative tasks, more consistent with cancer cells, which is out of scope of this work.

The other two groups, as one would expect, are clustered by condition, healthy and cancer cells, which shows that there are significant differences between both types of models, regardless of type of data and algorithm. Within each of these two clusters, two sub-clusters emerge, one for each data source (GEB and HPA), showing that the data source seems to be more discriminant than the algorithm.

Finally, within these sub-clusters, including three models from three algorithms, FASTCORE models are always closer to mCADRE, with MBA further apart. This is expected given the way the different algorithms are designed, as explained above.

Another test to the models obtained was conducted by performing an enrichment analysis. With the objective of evaluating the processes that are lost and gained by the cancer cells when compared to the normal ones, a p value of 0.025 was used while using the *Category* and *GOstats* packages from *Bioconductor*. With the proteomics data as input, processes contributing to the production of small molecules (e.g., nucleotides) were lost in the MBA models, while the pathways related to the metabolism of fatty acids were enriched. The processes lost in the mCADRE model were similar to the MBA one, while ion transport processes

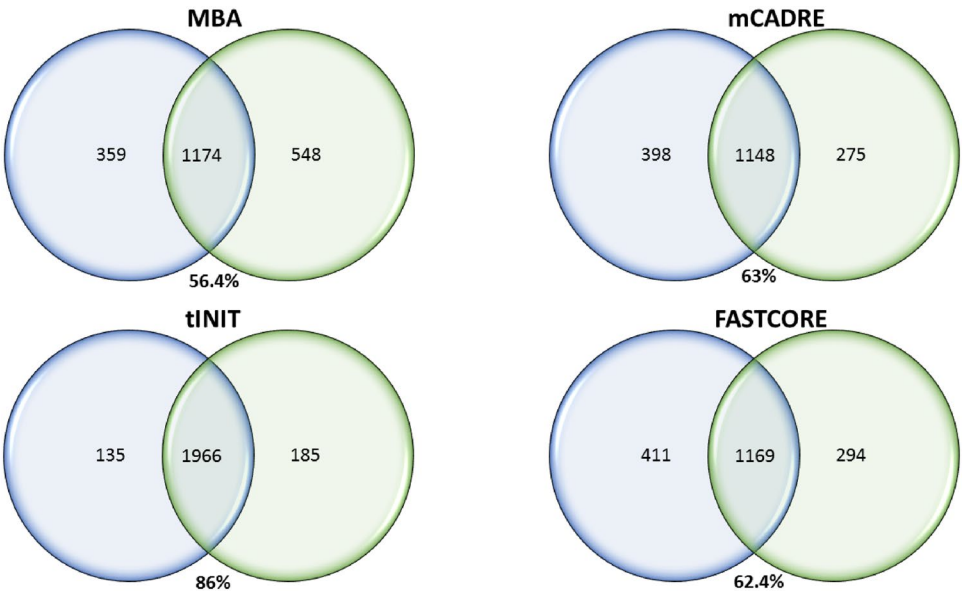

Fig. 1 Common and exclusive reactions between Normal and Cancer cells from HPA data using MBA, mCADRE, tINIT and FASTCORE algorithms. The values under each *Venn diagram* represent the percentage of shared reactions for both conditions. The *blue one* is relative to the normal model and the *green* to the cancer one

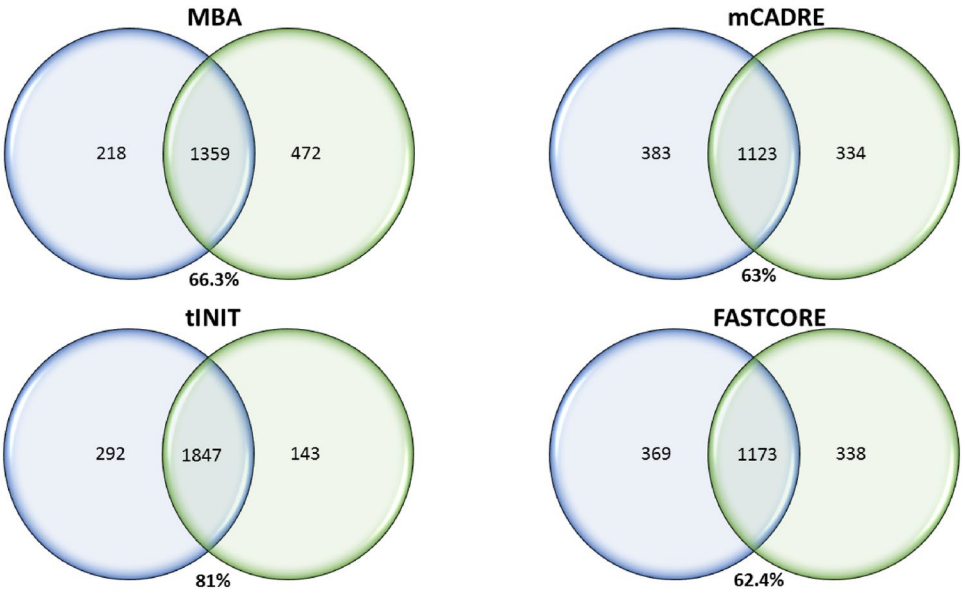

Fig. 2 Common and exclusive in reactions between normal and cancer cells from GEB data using MBA, mCADRE, tINIT and FASTCORE algorithms. The values under each *Venn diagram* represent the percentage of shared reactions for both conditions. The *blue one* is relative to the normal model and the *green* to the cancer one

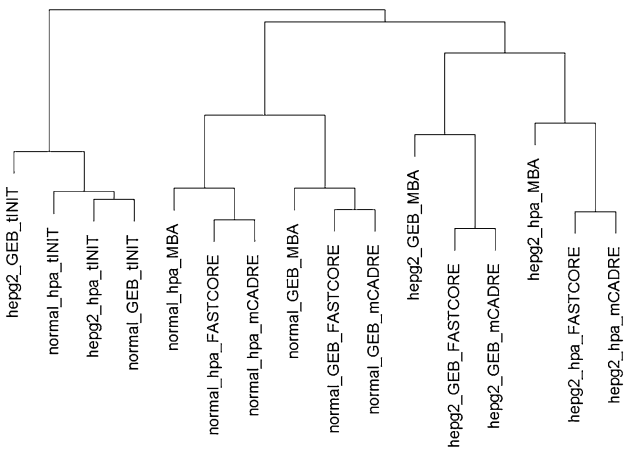

Fig. 3 Hierarchical clustering of all the 16 models generated with the method "complete"

were "acquired". In the tINIT one, processes lost were associated with the metabolism of fatty acids and enrichment with the production of ATP and nucleotides. Finally, for the FASTCORE the HepG2 model has an increase of production of ATP processes (like the tINIT) and small molecules resorting to different pathways.

For the transcriptomics data, MBA models lost mostly the same processes as the ones considering proteomics data, and the ones enriched were related to the metabolism of carboxylic and organic acids, and also to oxoacid processes. For the mCADRE models, the results were also similar to the MBA ones, with the addition of oxidation–reduction and ATP synthesis in mitochondria processes, while the processes that were enriched were also related to ions transport (in concordance with the proteomics data). Looking at the tINIT models, the processes lost were the same as the enriched ones on the MBA models for the transcriptomics data (metabolism of carboxylic and organic acids and oxoacid processes) and the ones enriched were the same as the ones enriched on the proteomics data by the same algorithm, production of ATP and nucleotides. At last, the FASTCORE algorithm shows lost of production of small molecules and acid metabolism and demonstrates an enrichment of same processes. This may suggest that the HepG2 cell line is obtaining the same metabolites through different reactions/pathways.

In another analysis, we decided to evaluate the performance of the models by verifying how many liver-specific metabolic tasks (from [9]) they could complete. From a total of 408 tasks tested, Recon1 can perform 281. Table 2 illustrates the percentage of tasks that our tissue-specific models can perform and the heatmap present in Fig. 4 shows which subset of the metabolic tasks are performed by each model (some subsets were removed since no models were able to perform any task).

There are several aspects in this analysis. Looking at the Table 2, we can see that the algorithm that has a higher percentage of tasks performed is the tINIT and in average the tumor models are able to fulfill around 5% more tasks than the normal ones.

Looking particularly at both models from the HPA data, they mainly differ in two aspects: the normal tissue-specific model is not capable of catabolizing bilirubin and biosynthesizing fatty acids; on the other hand, the cancer model is not able to biosynthesize creatine. It has been reported that a low level of production of creatine is common in liver cancer patients [43].

Looking at the GEB models, both are not capable of catabolizing bilirubin and transforming fatty acid (which at least the normal tissue should be able to accomplish), the tumor one is not capable of performing detoxification of xenobiotics. Indeed, it has been reported that the way these compounds are metabolized can affect the outcome of the liver cancer [44]. However, the tumor model is not able to perform gluconeogenesis and this is different from expected, since one of the treatments applied to this type of cancer is the inhibition of this pathway [45].

The algorithm that performed worse was without any doubt the mCADRE one. Even though the best model is the one based on the GEB data for the normal tissue, it is only capable of performing 26.7% of the tasks. Looking at the FASTCORE algorithm, none of them is capable of detoxification of xenobiotics, catabolism of bilirubin, fatty acid transformation or gluconeogenesis. Curiously, both HepG2 models for the FASTCORE algorithm are not able to biosynthesize sphingolipids.

Finally, looking at the MBA models, we can see that the model generated with the HPA data for the normal tissue is the "worst" model of the group. We can also verify that none of the tissue-specific models is capable of performing glycogenesis or gluconeogenesis or even fatty acid transformation.

As the final part of the work, we decided to "force" our cancer models to produce biomass. This makes biological sense, since cancer cells evolve to replicate as fast as possible; therefore, it is expected that they possess the cellular machinery to obtain all needed precursors.

The Recon1 model does not possess a biomass reaction. We, thus, retrieved this reaction from the Recon2 model [6] and introduced it to the Recon1 model. However, none of our tissue-specific models were capable of producing it. The Table 3 shows how many biomass precursors each model was able to achieve.

This analysis was achieved by performing an FBA in which the objective function was the maximization of the excretion of each metabolite and using the RPMI-1640 medium from Folger et al. [27]. tINIT and MBA have the best overall results for the production of the precursors of biomass, with mCADRE being the "least" capable of such task.

Due to this, we decided to add the reactions necessary for each model to fulfill the production of biomass. Table 4 shows the number of reactions needed to add to each model to be able to produce biomass.

As expected, the tINIT algorithm models are the ones who need the least number of reactions to be able to produce biomass. MBA reconstructed models need more reactions to be introduced into their models. Again, the mCADRE algorithm showed the highest number of reactions needed. It is also worth noticing that in the general case, the tumor models produce more precursors and need less reactions to be able to produce biomass, which may be biologically plausible.

As the final objective of this work, we decided to perform an FBA to evaluate the differences in the production of biomass for the different cancer models generated (Table 5).

Since the production of the Recon1 model with the biomass reaction is also 0.084 mmol.gDW^{-1} h^{-1}, there are four models which can achieve the same amount of biomass production and mCADRE has the lowest overall amount. This shows that the generated cancer models are able to grow at the maximum theoretical level, which is the one defined by the template global GSMM.

Table 2 Percentage of performed tasks by condition and algorithm of the 281 tasks that Recon1 is able to perform

	MBA (%)	mCADRE (%)	tINIT (%)	FASTCORE (%)	Mean (%)
Normal_HPA	7.8	9.6	87.9	58.4	50.7
Normal_GEB	55.2	26.7	88.6	71.5	
HepG2_HPA	63	2.5	92.5	40.5	55.3
HepG2_GEB	79.4	10.3	76.5	71.9	

4 Conclusions and Future Perspectives

The full understanding of the mechanism that lies beyond a human cell is still far away. Different layers of cell functions like metabolism or regulation are still not fully understood, making a difficult task to merge all the knowledge that we currently have. However, the integration of transcriptomics/proteomics data to infer the metabolism of a certain type of cell can give us some important insights.

Using the workflow presented in this work, we were able to evaluate how different algorithms can be used for two types of data and reconstruct a tissue-specific model for the healthy and cancer liver cells. As most of other cancer types, liver cancer is a disease which kills millions of humans and for which we still have few knowledge. With this kind of approach, in spite of many limitations, we are able to simulate what occurs in its metabolism and study the differences between the different approaches can lead to a panoply of results.

In this particular study, the algorithm that yielded more consistent and probably biological meaningful results was the tINIT. It was the one closer to simulate

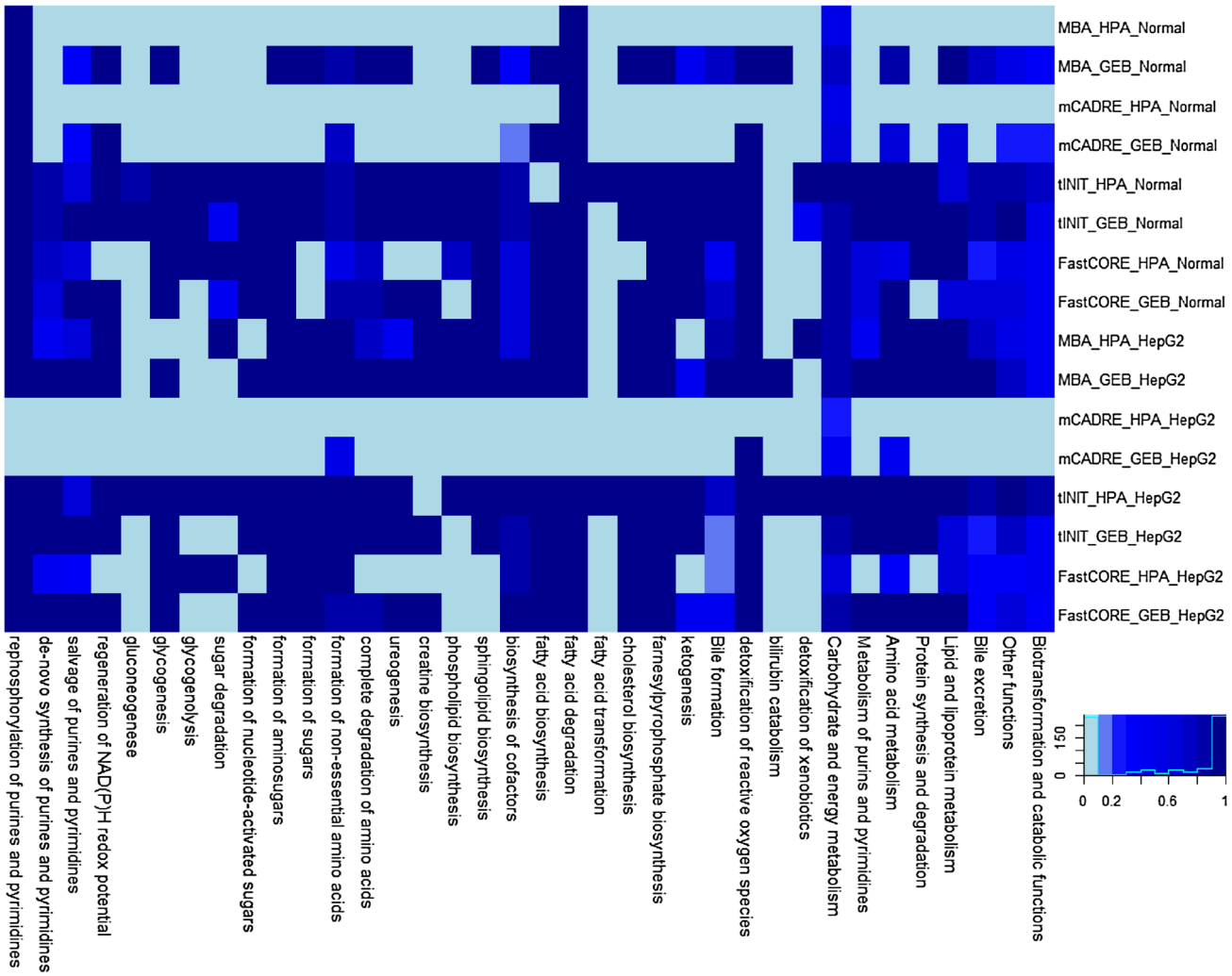

Fig. 4 Heatmap illustrating the percentage of the subset of the metabolic tasks (that can be done by Recon1) performed by each tissue-specific metabolic models

Table 3 Number of precursors that each cancer model had before the inclusion of the new reactions to be able to produce biomass (38 in total)

	MBA	mCADRE	tINIT	FASTCORE
HepG2_HPA	29	9	32	23
HepG2_GEB	33	14	28	26

Table 4 Number of reactions added to each cancer model to be able to produce biomass

	MBA	mCADRE	tINIT	FASTCORE
HepG2_HPA	17	28	8	26
HepG2_GEB	9	30	10	16

Table 5 Production of biomass by the cancer models after the integration of the necessary reactions (in mmol $gDW^{-1}\,h^{-1}$)

	MBA	mCADRE	tINIT	FASTCORE
HepG2_HPA	0.084	0.003	0.069	0.029
HepG2_GEB	0.084	0.012	0.084	0.084

a liver cell and the differences in the performance of the tasks could lead to a better understanding of how the metabolism could be an important target for the therapy of the liver cancer. However, the models generated by this algorithm are the ones that comprise more reactions, which can imply that they are "closer" to the Recon1 and possibly not being as informative as desired. In addition, if we look at the clustering of the models, the groups formed by the tINIT cluster together normal and cancer

models, which was not expected and may lead to "false positives".

With this in mind, the next choice would be the MBA algorithm. Even if the FASTCORE algorithm has fewer reactions, FASTCORE clusters closer to the mCADRE models, which performed poorly in the tasks and production of biomass. MBA models contained more precursors and needed less reactions to be added to the model to be able to produce biomass and both models produced the same amount as the Recon1 model. The main problem with this algorithm is the need to create a good number of "submodels" which is time consuming (in this aspect, the FASTCORE algorithm is much faster).

All algorithms have their advantages and disadvantages, so it is not easy to pick the "ideal" algorithm. The use of other template models, the creation of new algorithms for tissue-specific reconstructions and the integration of other types of data, like regulation of gene expression, could improve the knowledge that we have for this specific case and other important case studies.

Acknowledgements This study was supported by the Portuguese Foundation for Science and Technology (FCT) under the scope of the strategic funding of UID/BIO/04469/2013 unit and COMPETE 2020 (POCI-01-0145-FEDER-006684), BioTecNorte operation (NORTE-01-0145-FEDER-000004) and Search-ON2: Revitalization of HPC infrastructure of UMinho, (NORTE-07-0162-FEDER-000086), all funded by European Regional Development Fund under the scope of Norte2020—Programa Operacional Regional do Norte.

References

1. Price ND, Reed JL, Palsson BØ (2004) Genome-scale models of microbial cells: evaluating the consequences of constraints. Nat Rev Microbiol 2(11):886–897
2. Orth JD, Thiele I, Palsson BØ (2010) What is flux balance analysis? Nat Biotechnol 28(3):245–248
3. Ryu JY, Kim HU, Lee SY (2015) Reconstruction of genome-scale human metabolic models using omics data. Integr Biol 7(8):859–868
4. Duarte N, Becker SA (2007) Global reconstruction of the human metabolic network based on genomic and bibliomic data. Proc Natl Acad Sci USA 104(6):1777–1782
5. Tortora GJ, Derrickson BH (2012) Principles of anatomy and physiology. Wiley, Hoboken
6. Thiele I, Swainston N, Fleming RM, Hoppe A, Sahoo S, Aurich MK, Haraldsdottir H, Mo ML, Rolfsson O, Stobbe MD et al (2013) A community-driven global reconstruction of human metabolism. Nat Biotech 31(5):419–425
7. Mardinoglu A, Agren R, Kampf C, Asplund A, Uhlen M, Nielsen J (2014) Genome-scale metabolic modelling of hepatocytes reveals serine deficiency in patients with non-alcoholic fatty liver disease. Nat Commun 5:3083
8. Hao T, Ma H-W, Zhao X-M, Goryanin I (2010) Compartmentalization of the Edinburgh Human Metabolic Network. BMC Bioinform 11:393
9. Gille C, Bölling C, Hoppe A, Bulik S, Hoffmann S, Hübner K, Karlstädt A, Ganeshan R, König M, Rother K et al (2010) Hepatonet1: a comprehensive metabolic reconstruction of the human hepatocyte for the analysis of liver physiology. Mol Syst Biol 6(1)
10. Sahoo S, Franzson L, Jonsson JJ, Thiele I (2012) A compendium of inborn errors of metabolism mapped onto the human metabolic network. Mol BioSyst 8(10):2545
11. Sahoo S, Thiele I (2013) Predicting the impact of diet and enzymopathies on human small intestinal epithelial cells. Hum Mol Genet 22(13):2705–2722
12. Agren R, Bordel S, Mardinoglu A, Pornputtapong N, Nookaew I, Nielsen J (2012) Reconstruction of genome-scale active metabolic networks for 69 human cell types and 16 cancer types using init. PLoS Comput Biol 8(5):e1002518
13. Mardinoglu A, Agren R, Kampf C, Asplund A, Nookaew I, Jacobson P, Walley AJ, Froguel P, Carlsson LM, Uhlen M, Nielsen J (2013) Integration of clinical data with a genome-scale metabolic model of the human adipocyte. Mol Syst Biol 9:649
14. Kanehisa M, Goto S, Sato Y, Furumichi M, Tanabe M (2012) KEGG for integration and interpretation of large-scale molecular data sets. Nucleic Acids Res 40(D1)
15. Romero P, Wagg J, Green ML, Kaiser D, Krummenacker M, Karp PD (2005) Computational prediction of human metabolic pathways from the complete human genome. Genome Biol 6(1):R2
16. Fahy E, Subramaniam S, Murphy RC, Nishijima M, Raetz CR, Shimizu T, Spener F, van Meer G, Wakelam MJ, Dennis EA (2009) Update of the LIPID MAPS comprehensive classification system for lipids. J Lipid Res 50(Suppl):S9–14
17. Croft D (2013) Building models using reactome pathways as templates. Methods Mol Biol 1021:273–283
18. Palsson B (2002) In silico biology through "omics". Nat Biotechnol 20(7):649–650
19. Lewis NE, Nagarajan H, Palsson BØ (2012) Constraining the metabolic genotype-phenotype relationship using a phylogeny of in silico methods. Nat Rev Microbiol 10(4):291–305
20. Hyduke DR, Lewis NE, Palsson BØ (2013) Analysis of omics data with genome-scale models of metabolism. Mol BioSyst 9(2):167–174
21. Hoppe A (2012) What mRNA abundances can tell us about metabolism. Metabolites 2(4):614–631
22. Palsson B, Zengler K (2010) The challenges of integrating multi-omic data sets. Nat Chem Biol 6(11):787–789
23. Covert MW, Knight EM, Reed JL, Herrgard MJ, Palsson BO (2004) Integrating high-throughput and computational data elucidates bacterial networks. Nature 429(6987):92–96
24. Orth JD, Thiele I, Palsson BØ (2010) What is flux balance analysis? Nat Biotech 28(3):245–248
25. Park J, Park H, Kim W, Kim H, Kim T, Lee S (2012) Flux variability scanning based on enforced objective flux for identifying gene amplification targets. BMC Syst Biol 6(1):106
26. Kim HU, Kim SY, Jeong H, Kim TY, Kim JJ, Choy HE, Yi KY, Rhee JH, Lee SY (2014) Integrative genome-scale metabolic analysis of Vibrio vulnificus for drug targeting and discovery. Mol Syst Biol 7(1):460–460
27. Folger O, Jerby L, Frezza C, Gottlieb E, Ruppin E, Shlomi T (2011) Predicting selective drug targets in cancer through metabolic networks. Mol Syst Biol 7(1)
28. Frezza C, Zheng L, Folger O, Rajagopalan KN, MacKenzie ED, Jerby L, Micaroni M, Chaneton B, Adam J, Hedley A et al (2011) Haem oxygenase is synthetically lethal with the tumour suppressor fumarate hydratase. Nature 477(7363):225–228
29. Jerby L, Wolf L, Denkert C, Stein GY, Hilvo M, Oresic M, Geiger T, Ruppin E (2012) Metabolic associations of reduced proliferation and oxidative stress in advanced breast cancer. Cancer Res 72(22):5712–5720
30. Mardinoglu A, Gatto F, Nielsen J (2013) Genome-scale modeling of human metabolism—a systems biology approach

31. Baffy G, Brunt EM, Caldwell SH (2012) Hepatocellular carcinoma in non-alcoholic fatty liver disease: an emerging menace. J Hepatol 56(6):1384–1391
32. Jemal A, Bray F, Center MM, Ferlay J, Ward E, Forman D (2011) Global cancer statistics. CA 61(2):69–90
33. Kampf C, Mardinoglu A, Fagerberg L, Hallström BM, Edlund K, Lundberg E, Pontén F, Nielsen J, Uhlen M (2014) The human liver-specific proteome defined by transcriptomics and antibody-based profiling. FASEB J 28(7):2901–2914
34. Jerby L, Shlomi T, Ruppin E (2010) Computational reconstruction of tissue-specific metabolic models: application to human liver metabolism. Mol Syst Biolo 6(1)
35. Wang Y, Eddy JA, Price ND (2012) Reconstruction of genome-scale metabolic models for 126 human tissues using mcadre. BMC Syst Biol 6(1):153
36. Correia S, Rocha M (2015) A critical evaluation of methods for the reconstruction of tissue-specific models. In: Proc. 17th Portuguese Conference on Artificial Intelligence, EPIA 2015, Coimbra, Sep 8–11, 2015, pp 340–352
37. Knowles BB, Howe CC, Aden DP (1980) Human hepatocellular carcinoma cell lines secrete the major plasma proteins and hepatitis b surface antigen. Science 209(4455):497–499
38. Duarte NC, Becker SA, Jamshidi N, Thiele I, Mo ML, Vo TD, Srivas R, Palsson BØ (2007) Global reconstruction of the human metabolic network based on genomic and bibliomic data. PNAS 104(6):1777–1782
39. Uhlen M, Oksvold P, Fagerberg L, Lundberg E, Jonasson K, Forsberg M, Zwahlen M, Kampf C, Wester K, Hober S et al (2010) Towards a knowledge-based human protein atlas. Nature Biotech 28(12):1248–1250
40. McCall MN, Jaffee HA, Zelisko SJ, Sinha N, Hooiveld G, Irizarry RA, Zilliox MJ (2014) The gene expression barcode 3.0: improved data processing and mining tools. Nucleic Acids Res 42(D1):D938–D943
41. Agren R, Mardinoglu A, Asplund A, Kampf C, Uhlen M, Nielsen J (2014) Identification of anticancer drugs for hepatocellular carcinoma through personalized genome-scale metabolic modeling. Mol Syst Biol 10(3)
42. Vlassis N, Pacheco MP, Sauter T (2014) Fast reconstruction of compact context-specific metabolic network models. PLoS Comput Biol 10(1)
43. Chen J, Wang W, Lv S, Yin P, Zhao X, Lu X, Zhang F, Xu G (2009) Metabonomics study of liver cancer based on ultra performance liquid chromatography coupled to mass spectrometry with hilic and rplc separations. Anal Chim Acta 650(1):3–9
44. Williams GM (1980) The pathogenesis of rat liver cancer caused by chemical carcinogens. Biochim Biophys Acta (BBA) Rev Cancer 605(2):167–189
45. Wang B, Hsu S-H, Frankel W, Ghoshal K, Jacob ST (2012) Stat3-mediated activation of microrna-23a suppresses gluconeogenesis in hepatocellular carcinoma by down-regulating glucose-6-phosphatase and peroxisome proliferator-activated receptor gamma, coactivator 1 alpha. Hepatology 56(1):186–197